Interactive Reasoning: Visualizing and Controlling Chain-of-Thought Reasoning in Large Language Models

Rock Yuren Pang
University of Washington
Seattle, Washington

K. J. Kevin Feng
University of Washington
Seattle, Washington

Shangbin Feng
University of Washington
Seattle, Washington

Chu Li
University of Washington
Seattle, Washington

Weijia Shi
University of Washington
Seattle, Washington

Yulia Tsvetkov
University of Washington
Seattle, Washington

Jeffrey Heer
University of Washington
Seattle, Washington

Katharina Reinecke
University of Washington
Seattle, Washington

Abstract

The output quality of large language models (LLMs) can be improved via "reasoning": generating segments of chain-of-thought (CoT) content to further condition the model prior to producing user-facing output. While these chains contain valuable information, they are verbose and lack explicit organization, making them tedious to review. Moreover, they lack opportunities for user feedback, such as to remove unwanted considerations, add desired ones, or clarify unclear assumptions. We introduce *Interactive Reasoning*, an interaction design that visualizes chain-of-thought outputs as a hierarchy of topics and enables user review and modification. We implement interactive reasoning in Hippo, a prototype for AI-assisted decision making in the face of uncertain trade-offs. In a user study with 16 participants, we find that interactive reasoning in Hippo allows users to quickly identify and interrupt erroneous generations, efficiently steer the model towards customized responses, and better understand both model reasoning and model outputs. Our work contributes to a new paradigm that incorporates user oversight into LLM reasoning processes.

CCS Concepts

• **Human-centered computing** → **User interface design**; **Empirical studies in HCI**.

Keywords

Interactive Reasoning, User Experience, Interaction Design

1 Introduction

There has been a surge of interest in developing and studying the reasoning capabilities of large language models (LLMs) [58, 92]. So-called *reasoning models*, such as OpenAI's `o3` [24] and DeepSeek's `R1` [16] models, include a "reasoning step" before generating their response and can perform complex tasks, align with social values, and adapt to user preferences [24, 66]. The reasoning step is often referred to as *test-time scaling* [49, 83], where a model is allowed to allocate significantly more computational resources during the inference phase to improve reasoning ability. Due to recent advances in LLMs, reasoning models have become more efficient and less costly to develop and use [16, 49].

While these reasoning steps are seen as a positive development for transparent LLMs [3], users have little control over the reasoning steps to shape the model's reasoning process. In addition, the reasoning step from test-time scaling is generally verbose and unstructured, making it tedious for users to make sense of the reasoning. For users, making sense of and having control over the reasoning is especially important when seeking advice from models in high-stake domains such as ethical, financial, medical decision-making, where reasoning steps may be misaligned with a user's core beliefs, knowledge, and priorities [20]. For example, users' values and perspectives should ideally inform the reasoning process in a decision-making process [13]. When users identify misalignments in LLM reasoning, they cannot provide targeted feedback when the model makes incorrect assumptions. Users must work in a cycle where they issue a new prompt, review the model output, review reasoning steps, and manually refine their output.

In this paper, we introduce *Interactive Reasoning* to reimagine how users engage with LLMs' reasoning processes. Our work was inspired by the longstanding challenge of balancing automation and human control in AI systems [19, 64], as well as recent work on sensemaking [25, 70] and appropriate AI reliance in the context of generative AI [6, 32]. Our approach transforms complex reasoning chains into interactive tree representations, which enables users to visualize, directly edit reasoning steps, provide feedback, and shape the model's final output. By using an interactive tree representation, our approach uplifts and scaffolds the intermediate steps of a reasoning model, shows (dis)connections to the model output, and enables interactions with tree nodes to steer the model output.

We instantiate *Interactive Reasoning* in Hippo[1], a prototype that allows users directly interact with the reasoning process before the model generates its final output. In a user study with 16 participants, we used Hippo to compare with an editable baseline interface and to answer three key research questions around (1) users' control, sense-making, and awareness when interacting with the reasoning chains; (2) users' perception over the final response after acting with the reasoning chains; and (3) users' interactions with the interactive reasoning tree. We additionally demonstrate Hippo's use cases where users interact with reasoning chains in diverse

[1]We named our system after the 5th-century thinker Augustine of Hippo.

tasks outside of our study context such as information seeking and financial planning. To summarize, our work contributes:

(1) A novel interaction design with reasoning models, which we instantiate through Hippo, a research prototype to visualize the reasoning steps, allow direct human feedback, and show (dis)connections between the reasoning steps and the model output.
(2) Empirical findings from a controlled user study showing that participants indicated significantly more control over, sense-making, and awareness of assumptions in the reasoning when using Hippo compared to using a baseline interface. Participants reported an increased confidence in making the final decision with Hippo compared to the baseline. Qualitative results showed that participants valued the transparency of viewing the reasoning chains and the ability to repurpose the model response after engaging with the reasoning chain.

2 Related Work

2.1 What is Reasoning?

Much of the definition of reasoning sprang out of Aristotle's theory of the syllogism, where he defined reasoning (*syllogismos*). It generally refers to the making of assumptions called *premises* and the process of moving toward conclusions (*end point*) from these assumptions by rules [74]. To conceptualize reasoning, logicians consider that reasoning can be modeled abstractly by a graph or argument diagram where arcs (steps) link points (vertices in the graph) [75]. In logic, laying out an argument structure is also referred to as "argument diagramming", which aims to transfer arguments into a structured representation to evaluate them [55, 59].

In NLP, "reasoning" is a process of answering questions that require complex, multi-step generation with intermediate steps [58], though current LLMs are still not capable of genuine logical reasoning [48]. In this process, the reasoning model integrates multiple knowledge (e.g., encyclopedic and commonsense knowledge) to derive some new conclusions about the (realistic or hypothetical) world [88]. Knowledge can be derived from sources that are both explicit and implicit. Conclusions are assertions or events assumed to be true in the world, or practical actions [89]. There are a number of ways to build and improve reasoning models, such as tree-based search methods [41, 66, 83, 85] and reinforcement learning [16]. Recent methods to improve test-time scaling emphasized improving the final model output (rather than on the length or steerability in our paper), leveraging methods such as Monte-Carlo Tree Search (MCTS) [84], process reward models [83], and budget forcing [49].

Our work builds on this foundation by modeling the reasoning processes of LLMs as graphs. Unlike previous research that focuses on reasoning that follows a defined argumentative structure, e.g., essays [69], debates [14], and political rhetoric [67], we emphasize capturing the intermediate reasoning steps from a simple query to LLMs. These steps, which are typically internal to the model, are made visible to users, enabling transparency and interaction with the reasoning process. This transparency not only allows users to see the various topics the reasoning includes, and how the final response is derived but also enables interaction with the model reasoning process during test-time scaling.

2.2 Appropriate AI Reliance

With the increasing LLM capabilities, users increasingly seek guidance from LLMs for decision-making in daily life [11, 52, 94]. These decision-making processes are not clear-cut and depend on trade-offs on the contexts, personal values, and ethical standards [17]. However, the prevalence of LLMs has raised questions about over-reliance and LLMs' impact on critical thinking skills and practices when making such personal decisions [38]. Prior work in explainable AI (XAI) has studied users' appropriate reliance on AI extensively. The concept of appropriate reliance can typically be defined as "relying on the AI when it's correct, and relying on yourself when it's not" [61]. To address overreliance, prior work has investigated solutions, e.g., providing information about the AI's performance [87], explanation of outputs [5], and communication of uncertainty [93]. A notable example is the cognitive forcing function—interventions being applied at the decision-making time to disrupt heuristic reasoning and thus cause the person to engage in analytical thinking [8].

Recent work has built on these work and turned to how end users can appropriately rely on LLMs and incorporate LLMs into their decision-making process [6, 31]. These work mostly focus on user reliance in the context of answering objective questions (e.g., facts, LSAT questions). However, many daily decisions do not have a clear-cut answer and arguably require users to make trade-offs based on their unique contexts. Our work builds on strategies from the XAI community to foster appropriate reliance on AI, and redesign the intermediate reasoning process to explore how users may engage with decisions with the long reasonig chain.

2.3 Diagramming for Large Language Models

HCI has contributed systems to explore and evaluate LLMs' output via diagrams. For example, Graphologue [25] and Sensecape [70] help users interact non-linearly on an interface to help users understand and explore LLM-generated information in a node-link diagram. In a similar vein, prior work have leveraged the node-link diagrams to explore different topical aspects in, e.g., data analysis [44], creative coding [1], responsible AI [79], research ideation [57]. These tools, as Arawjo et al. [2] dubbed, are sensemaking interfaces for information foraging. Another thread of novel visual interaction of LLMs include systems for designing LLM-based applications, such as PromptChainer[81], which construct "AI chains" [82], or data feeds between LLM and other tools or scripts.

In addition to tools to show the flow of the LLM-generated information, there is a large number of work focusing on interactive evaluation for LLM prompts and output [53]. For instance, ChainForge is a visual toolkit for prompt engineering and on-demand hypothesis testing of text generation of LLMs [2]. EvalLM is an interface that aids users in revising prompts with synthetic LLM-based evaluators by providing the difference of the outputs [34]. LLM Comparator analyzes the LLM results from automatic side-by-side evaluation, thereby allowing users to understand when and why a model performs better or worse than a baseline model, and how the responses from two models are qualitatively different [28].

Our work focus on the reasoning process at test-time scaling, where users can provide feedback. In doing so, users also make sense of the reasoning steps, and enable the linkage between the reasoning

and the output. Our goal is to allow users to directly scrutinize the model reasoning, and explore the design opportunities for human participation in reasoning during test-time scaling.

3 Design Goals

We formulate our concrete design goals (DG) by drawing upon research in the HCI, UIST, and XAI communities around interacting and sensemaking with LLMs, as well as on avoiding AI overreliance.

DG1: Allow users to directly manipulate reasoning chains. While several computational approaches have been developed to address the challenge of aligning LLMs with human values [68], these models may still struggle to automatically resolve value conflicts in complex real-world decision-making tasks that require human intervention for trade-offs [11]. Users frequently encounter outputs that are overly general responses that may not correctly reflect the user's context [35]. *Interactive Reasoning* should allow users to directly give feedback to the intermediate reasoning process of the model, in line with the principles of direct manipulation of LLMs [45, 63]. This avoids the usual trial-and-error (i.e., iterative prompting and excessive turn-taking with AI) that is required when trying to obtain a satisfactory result [12, 90].

DG2: Encourage users to cognitively engage with assumptions made in the reasoning chain. The reasoning chain can often be regarded as supporting details or justification for the LLM response [80]. Prior work in XAI suggests that in many settings, the very presence of an explanation can increase users' trust and reliance (which can also result in overreliance) [5, 77]. However, cognitive forcing functions [8] have been shown to compel people to engage more thoughtfully with AI-generated explanations and reduce over-reliance. Further, a waiting time before model output may help users gain useful insight and reflect on the task [54]. Reviewing the intermediate reasoning steps can potentially serve as an effective means of involving users in critically evaluating the model's reasoning steps, encouraging them to verify underlying assumptions rather than passively accepting results.

DG3: Use graphical representations to reduce the information load for long reasoning chains. While requiring users to *read* the entire reasoning chain carefully would be ideal to ensure transparency, doing so is impractical, especially under limited time [71]. *Interactive Reasoning* should support effective communication of ideas, especially in the often long reasoning chains such as DeepSeek-R1. In HCI and visualization, graphical representations have been used to support sensemaking [4], in many domains [57, 70, 79], including for LLMs. For example, Graphalogue [25], which leverages node-link diagrams generated from LLMs's final output, has been shown to help users quickly grasp key concepts and their connections. We draw on this idea to make sense of the long reasoning chains, rather than the final output as done in [25]. In addition, the graphical representation should be complete even if it means visualizing a large tree to ensure full transparency of the reasoning chain. Similar to the Wikum system's summarization feature [91], users should be able to prune the tree to shorten the reasoning steps.

DG4: Provide timely opportunities for users to intervene and steer the model via reasoning. In a similar vein to DG3, requiring users to *edit* every idea in a reasoning chain is unrealistic. *Interactive Reasoning* requires balanced control between automation and human agency [23]. When dealing with large texts or datasets, decomposing the information can be beneficial, but interacting with each component inevitably makes recalling the context difficult [30, 82]. From early feedback on our prototype, we observed that users quickly lost the patience to interact with all the reasoning components if the chain is too long. *Interactive Reasoning* requires users to intervene on smaller information components, and only when the model requires feedback.

DG5: Attribute the final output to specific parts of the reasoning chain. Prior work has found that the presence of sources impacts the credibility of the outputs of LLM-infused applications [32, 33]. When reading the final output of the LLM model in high-stakes decision-making tasks, understanding the statements and their relations to the assumptions in the reasoning chain explain how the final response is generated. This linkage, as prior work suggested, can improve users' sense of control and ownership of the text generation process [15, 22, 29, 30], and has been employed in many RAG-based models and platforms, e.g., Perplexity.

4 Interactive Reasoning

In this section, we describe *Interactive Reasoning*, an interaction design for breaking up a reasoning chain into smaller topical units, visualizing the units in a hierarchy, and editing units via user feedback. Our current Interactive Reasoning implementation includes four interactive operations: one can *add*, *edit*, or *delete* units, or *regenerate* the model's reasoning chain. We detail the implementation of *Interactive Reasoning* in an interface called Hippo.

4.1 Hippo User Interface

We detail the features of Hippo that support each design goal in Section 3. At its core, Hippo allows users' control of an LLM's reasoning chains displayed in a hierarchical tree structure (rather than a linear sequence of plain text), which can in turn steer the final model response. We use a prompt – *"Where should I travel to during the spring break?"* to walk through the interface in Figure 1.

Interactive Preorder Traversal Tree Playground. After clicking the "Ask" button, the interactive reasoning tree appears as shown in Figure 1. The tree progressively generates its nodes (B) in a pre-order (depth-first) tree traversal sequence, where the parent (topic) node appears first, followed by the left subtree. This interface parses the original long reasoning chain into smaller, manageable nodes to reduce the reader's cognitive load. In each node, the text tokens are displayed to the user in "real-time". This sequential progression allows users to observe the model's thinking as it unfolds, rather than seeing only the completed output. Users can stop the tree progression, and focus on a specific node.

Interactive Nodes. For each node, users can directly revise the reasoning text by clicking the edit button (B) (**DG1**). Users can also generate additional nodes (F) and subtrees under a parent node to steer the reasoning subtopics using a custom prompt (C) (**DG4**). Users can regenerate or redirectly edit this node if desired. Users can delete a node or a subtree entirely if they choose to ignore the (sub)-topics. Users can also revisit the tree and revise the content inside a tree node after the tree generation is complete. Altogether,

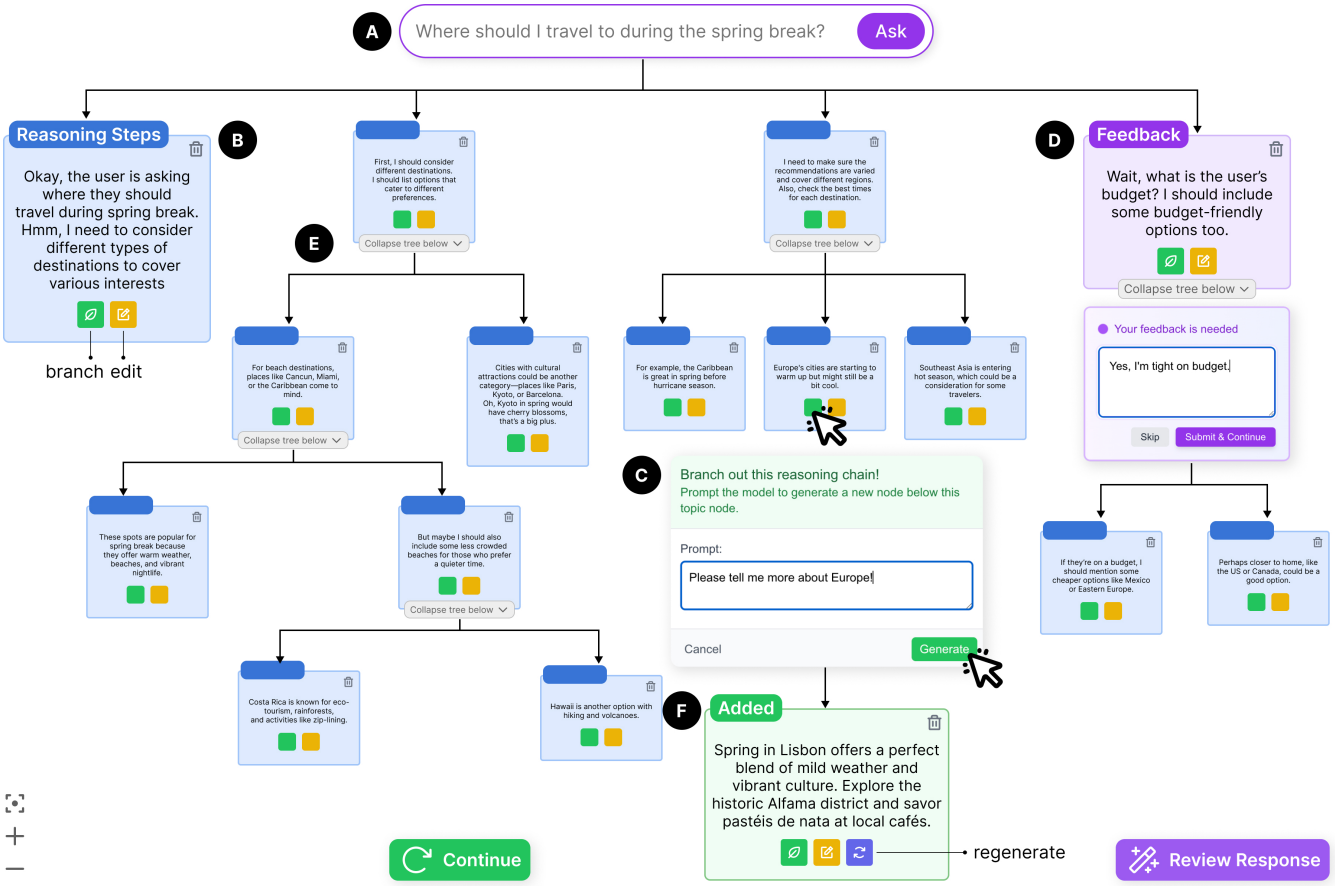

Figure 1: Hippo includes a tree visualization of the reasoning steps and allows users to directly control when models need users' feedback. Users input their query in the input bar A. Then, the reasoning tree progressively generates nodes B following a preorder (depth-first) tree traversal order. Users can branch out a reasoning node C by providing a customized prompt, which will add a new child node F. Hippo halts the tree generation at Feedback node to elicit feedback to clarify user contexts at D. Users can trim the tree by collapsing the subtree E, where Hippo append a summary of the subtree below a node. Users can pause and continue the generation at any point. Users organize the reasoning tree before reviewing the response in Figure 2.

the reasoning chain is completely represented in a tree structure (**DG3**), rather than a linear textual representation.

Hippo reasoning tree generation stops and prompts users to clarify the situation and give feedback D (**DG2**, **DG4**). Our early prototype asked users to stop at all given nodes to confirm the model reasoning, but this approach demanded excessive attention throughout the intermediate tree generation process. To address this, we implemented a Clarify step (Section 4.2.2) in the graph generation pipeline where only Feedback nodes that require clarification or context are surfaced to users. Users can choose to skip the answer. In the case of providing users' own feedback, Hippo automatically generates a follow-up node to users.

Interative Tree Trimming. When the full reasoning chain generation is completed, users can zoom out to view the tree overview and "trim" the tree by clicking E the "collapse the tree" button (**DG3**) to reduce the reasoning chain's size if it becomes too large. A summary of the collapsed subtree is appended below the immediate parent node. We acknowledge the inherent tradeoff between displaying the complete reasoning chain and abstracting information; however, our project's primary aim is to enhance the transparency of the reasoning process through a progressive tree-based visualization and to study users' perceptions after interacting with this information structure.

Visual Highlighting between Reasoning and Response. Users can edit the interactive reasoning tree and "review the response" based on the edited reasoning (Figure 2). Users can hover over the sentence in the response, and the Hippo highlights the connection between a sentence in the response H and the nodes in the reasoning tree G (**DG5**). Users can continue to edit the tree and update the response.

4.2 Tree Generation Pipeline

In *Interactive Reasoning* (Figure 3), the intermediate reasoning is decomposed using reasoning operators (describing the high-level

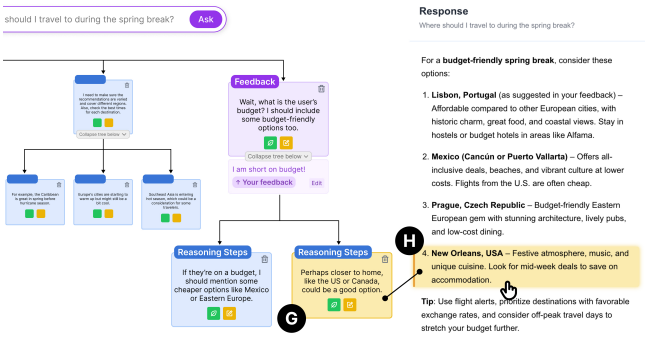

Figure 2: Hippo highlights the tree nodes G and sentences in the reasoning model final response H.

structure of reasoning) and tags (capturing the low-level details and important entities given structural constraints). This process requires structuring the raw text from the model's reasoning into a hierarchy of smaller components, drawing out components that require human intervention, and linking the final output to components in the reasoning chain.

4.2.1 Structure the Text. The Structure operator leverages the LLMs' capability to break down any unstructured text into topics [37]. The reasoning chain logic usually follows a *general-to-specific* deductive method of developing a topic [27, 50]. For instance, a paragraph may start with *"First, I should consider different types of travelers. I should list options that cater to different preferences"*. Then, this paragraph continues to dive into sub-topics: *"For beach destinations, places like Cancun, Miami, or the Caribbean come to mind."* After explaining these popular places, the reasoning chain further breaks down into another option within *"beach destination"* (e.g., *"But maybe I should also include some less crowded beaches for those who prefer a quieter time."*) The NLP community has demonstrated that LLMs have capabilities that organize the text into a hierarchy [26, 95]. Prior work has also shown that LLMs generally demonstrate capabilities in relationship awareness and structural understanding [9, 26, 40, 76], particularly using markups such as XML-like tags with few-shot prompting [16, 56, 65].

Building on these insights, our backend uses a few-shot prompt that instructs `GPT-4o` to identify the hierarchy of information. The instruction asks the model to annotate the original text inline with XML-like tags (i.e., `<topic>...</topic>` and `<branch>...</branch>`) to indicate the separation of text and hierarchy. In our pipeline, we do not rely on the paragraphs broken down by the original reasoning chain, as several paragraphs can discuss the same topic (this was also reflected in our pilot study and user evaluation, baseline condition). Instead, we aggregate text from the reasoning chain and segment it by topic 1 before applying the Structure operator 2 on these segments. We input smaller text chunks, rather than the monolithic reasoning text because LLM performance can degrade significantly with longer input length, making it difficult for a single long-context prompt to effectively cover all aspects [42]. Note that this component is different from prior work [25] in that we use the few-shot prompt to extract the *topic* hierarchy rather than important *entities* (e.g., nouns). We include our prompt template with an example output marked in the few-shot prompt in the supplementary material.

4.2.2 Flag nodes that need user feedback. Given a tagged concept hierarchy, the Clarify operator identifies the components that could benefit from human feedback. In the running example, a chunk of text such as "*Wait, what is the user's budget? I should include some budget-friendly options too.*" assumes that the user is looking for budget-friendly options. Users can quickly intervene and provide their budget expectations. For each node, we leverage the LLMs' ability to perform classification tasks, which can achieve accuracy comparable to human annotators [18, 47]. We used few-shot prompting to identify the `Feedback nodes` where user input would be valuable (e.g., uncertainty, preferences, personal experiences). To avoid latency (annotate node by node), we annotate the marked-up text with the additional tag `<user>...</user>`.

In practice, we found users became frustrated when presented with multiple similar questions from different branches of topics. Questions like "*How about the Caribbean?*" and "*Some spring breakers do cruises in the Caribbean. How about that?*" essentially ask the user to clarify the same question. To address this problem, we keep track of questions that have already been flagged and check for duplicates (using a cosine similarity measure on sentence vector in the embedding space with a `all-MiniLM-L6-v2` model threshold > 0.8) before showing new feedback nodes. We show our prompt for the Clarify operator in the supplementary material.

4.2.3 Generate a response based on the edited reasoning chain. When users provide feedback to questions or directly edit nodes in Hippo, we incorporate these contributions into the thought context. Once the user completes their edits, the pipeline updates the reasoning text enclosed within `<think>...</think>` tags [16]. We do not impose over-complicated system prompts in this step (e.g., imposing additional prompts other than the user's input to affect the final output). One goal of this paper is to examine if users find value in responses generated from the edited intermediate reasoning. We provide the prompt to elicit updated model response below (also see 4 in Figure 3).

```
{{original_text_prompt}}
<think>{{updated Thought Process}}</think>
<answer>[Updated Response to be generated]
```

4.2.4 Link the response to intermediate responses. The Link operator establishes connections between elements in the reasoning chain and corresponding segments in the final response. This functionality enables the traceability of how specific reasoning steps influence particular conclusions. We conceptualized this as a Natural Language Inference (NLI) task, where the reasoning segment serves as a premise and the response segment as a hypothesis. To achieve this, we prompt `GPT-4o` using a zero-shot prompting method to identify the connections [36].

We initially implemented this connection mechanism using the `bart-large` model [39] to evaluate semantic relationships between text segments. However, we encountered computational latency

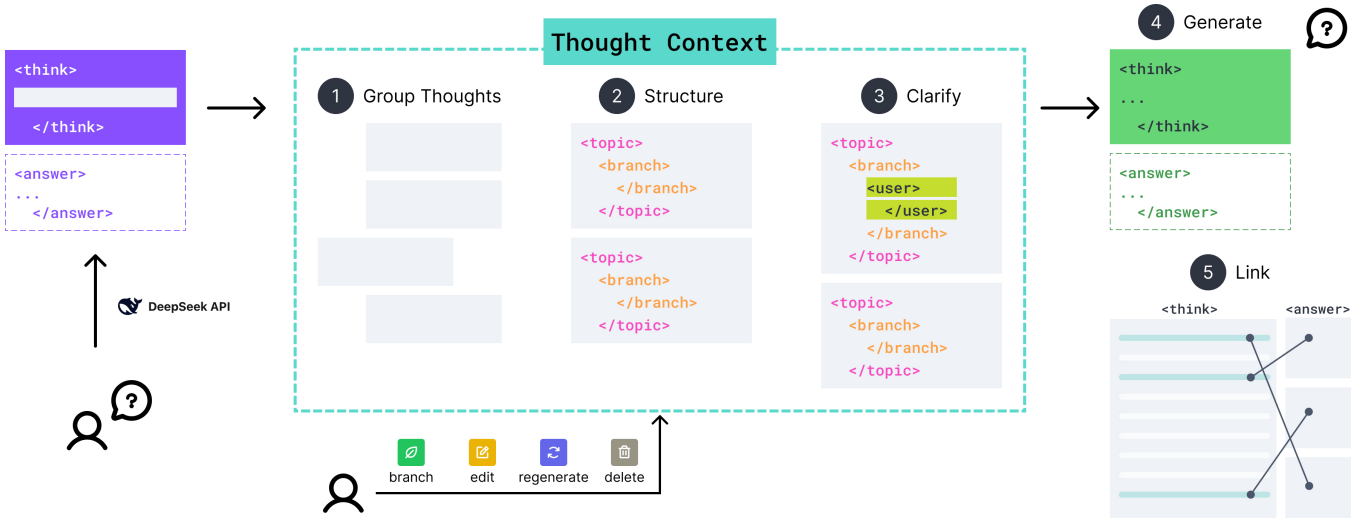

Figure 3: The *Interactive Reasoning* pipeline fetches the initial reasoning chain, structures the reasoning into topical hierarchy, flags text that might benefit from user intervention. The final output is directed back to the updated reasoning chain.

challenges in the case of long reasoning chains, which usually generate numerous reasoning nodes and response paragraphs. Processing times in batch for visual highlighting frequently exceeded one minute. In the end, we decided to use the the zero-shot prompting, as recent work has shown that LLMs exhibit strong zero-shot capabilities for NLI tasks without requiring task-specific fine-tuning [60, 78]. Additionally, prior work [21] showed that such models can effectively identify semantic relationships between text segments with performance comparable to specialized NLI models while offering significant computational advantages. We include the detailed prompt in the supplementary material.

4.3 Implementation

Hippo was implemented in the Next.js React framework and Tailwind CSS for styling, and a backend server using the Python Flask framework. We used `gpt-4o`, more specifically the `gpt-4o-2024-08-06`, for each of the reasoning operator, model to structure, clarify, and link the graphs. We used this for its fast API response time, computational efficiency, and low expense. The cost of `gpt-4o` was $2.50/1M tokens during the implementation. We used the open-weight `DeepSeek-R1` using the together.ai [73] as the reasoning model in the system implementation and the study.

5 User Evaluation

To evaluate Hippo, we conducted a within-subjects study with 16 participants. Each participant was asked to use a baseline system and the Hippo system. Our goal was to understand users' perception of *Interactive Reasoning* and explore design opportunities for future exploration. Concretely, our user evaluation was guided by the following three research questions.

- **RQ1**: How does Hippo compare to a baseline system in terms of users' sense of control over, sense-making of, and awareness of the reasoning steps?

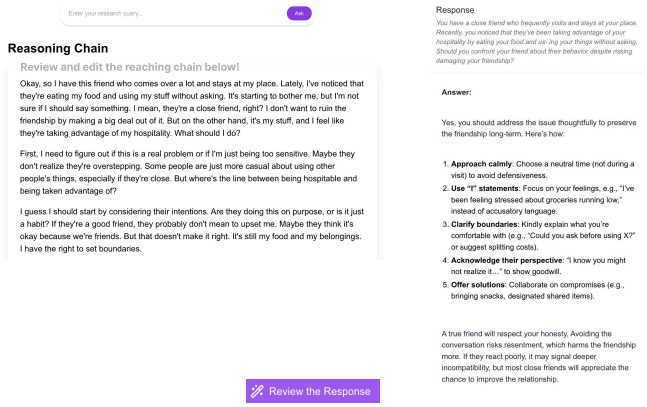

Figure 4: The baseline system. The reasoning steps show up progressively. Users can directly edit the reasoning steps in line before revealing the final response.

- **RQ2**: How do users perceive the final response after interacting with the reasoning steps?
- **RQ3**: How do users leverage interactive reasoning to understand and steer the model's behaviors?

5.1 Procedure

The study started with a brief introduction and overview of reasoning models. Then, participants were required to conduct tasks under two conditions: (1) a baseline system with regular reasoning output, which we adapted to allow editing as shown in Figure 4, and (2) the Hippo system that displays the reasoning steps in an interactive tree structure. We implemented the editable version as the baseline, as opposed to a non-editable version for a fair comparison. Comparing an editable interface to a non-editable one (e.g., on

DeepSeek or ChatGPT) would have likely yielded intuitive results for our RQs, especially on measure of the sense of control.

In each condition, participants asked a question selected from the DailyDilemma dataset [11], which represents relatable and realistic ethical dilemmas that require nuanced reasoning. We chose the daily dilemma situations to increase the stakes of the tasks in the controlled setting. Participants were told that they could modify the reasoning steps. We counterbalanced the prompts and conditions that each participant received to reduce the learning effects.

(1) Your parents are having a hard time managing their finances due to cognitive decline. They have always been fiercely independent and get extremely upset at the idea of you stepping in to help. Should you respect their independence and let them continue to struggle or intervene, potentially against their wishes, to ensure their financial stability?
(2) You have a close friend who frequently visits and stays at your place. Recently, you noticed that they've been taking advantage of your hospitality by eating your food and using your things without asking. Should you confront your friend about their behavior despite risking damaging your friendship?

After each condition, participants were directed to reveal the response by `DeepSeek-R1` and the updated response. Participants were then asked to fill out a short questionnaire, consisting of the following metrics (presented as seven-point Likert scale questions):

- **Control**: I could easily control the intermediate reasoning steps (e.g., by adding or revising the steps).
- **Sense-making**: I was able to make sense of the topics the LLMs navigated in their reasoning process.
- **Layout**: The layout of the information made it easy for me to understand the logic in the reasoning.
- **Awareness**: I was aware of the assumptions in the reasoning process.
- **Confidence**: By using the system, I feel confident about my potential decision in response to the prompt.
- **Insights**: I obtained useful insights from using the system.
- **Satisfaction**: I am satisfied with the response.
- **Caution**: I was fairly cautious when I read the response generated by the language model.
- **Future Use**: I could see myself integrating this system into my workflow when using AI to help with high-stakes decision-making.

The study concluded with a semi-structured interview that lasted around 15 minutes. Participants were asked to reflect on their experiences across the two interfaces and their previous experience with reasoning models such as `OpenAI-o1` and `DeepSeek-R1`. The interview was guided by questions about the usefulness and usability of both systems, as well as observations on how users interact with the reasoning tree during the study.

Analysis. We applied the Wilcoxon-Sign Rank test for the post-task Likert-scale questions. We conducted a thematic analysis of the semi-structured interviews. One author created an initial codebook from two interview sessions. Then, two authors came together to iteratively update the codebook in two sessions. The goal of the qualitative analysis was to identify emerging themes and challenges, rather than reach for strict inter-rater reliability [7].

5.2 Participants

To estimate the required number of participants, we performed an a-priori power analysis (Cohen's d = 0.8, $\alpha = 0.05$) and decided that a sample size of 16 participants is needed for detecting a medium effect between the baseline system and Hippo. We, therefore, recruited 16 participants through snowball sampling with the requirement that they had used LLM chatbots before. Our participants were 19-40 years of age, 9 male and 7 female, and had diverse backgrounds (from undergraduate students to working professionals in computer science, medicine, finance, education, and the movie industry). All reported using LLM chatbots more than once per week. Nine participants rarely reviewed the reasoning chain (i.e., 1-3 out of every 10 model runs), two never reviewed the reasoning chain, two sometimes reviewed them (i.e., around half of the time), two often reviewed reasoning (i.e., for a majority of model runs as part of a standard workflow), one always reviews the reasoning.

5.3 Results

In this section, we present quantitative and qualitative analyses of participants' data from our user study. We group these findings by our research questions. Participants' responses to the Likert-scale questions are shown in Figure 5.

5.3.1 Sense of Control, Sense-making, and Awareness of Assumptions (RQ1). First, participants appreciated the function of directly intervening in the model reasoning process. Based on the post-task ratings, there was a significant difference in perceived control between the baseline and Hippo ($M_{Baseline} = 4.19, M_{\text{Hippo}} = 5.75, p = 0.003$). The sense of control can be attributed to the `feedback` node where users need to provide their comments or clarification before moving forward in the reasoning chain, as well as the ability to directly add, edit, or regenerate nodes within the reasoning tree. In the study, we observed that all participants gave feedback to the nodes, and regenerated or added new content to the reasoning tree. P10 commented on their perceived ease of editing the reasoning steps: "*[Hippo]'s a direct manipulation of the reasoning, which is very nice, I see for me it was very easy to select and delete the whole thing that I didn't care about.*" P6 mentioned that they were *"empowered to take steps and make changes"*. Notably, while the baseline condition also supported users in directly editing the reasoning steps, participants commented that "*yeah, I know that I can edit, but I don't know where to start.*" [P4].

We also found that the graphical representation may improve their sense-making of the overall reasoning process. Participants rated Hippo higher than the baseline condition in sense-making of the long reasoning chain ($M_{Baseline} = 5.19, M_{\text{Hippo}} = 6.44, p = 0.004$) and attributed the improved sense-making to the different layout ($M_{Baseline} = 3.44, M_{\text{Hippo}} = 6.00, p = 0.009$). P8, a data analyst, commented "*That [Hippo] makes sense of the topics. That was one of the best parts of it, actually. It made so much sense to break down [like] the key ideas, and then it breaks down into more in-depth components.*" P9 stated that "*the representation made it easier to follow the reasoning, see different paths, and understand the differences between them more readily than with linear text.*"

In the study, we also asked users for their awareness of assumptions in the reasoning process. The post-task ratings indicate a significant difference in the awareness of assumptions ($M_{Baseline} =$

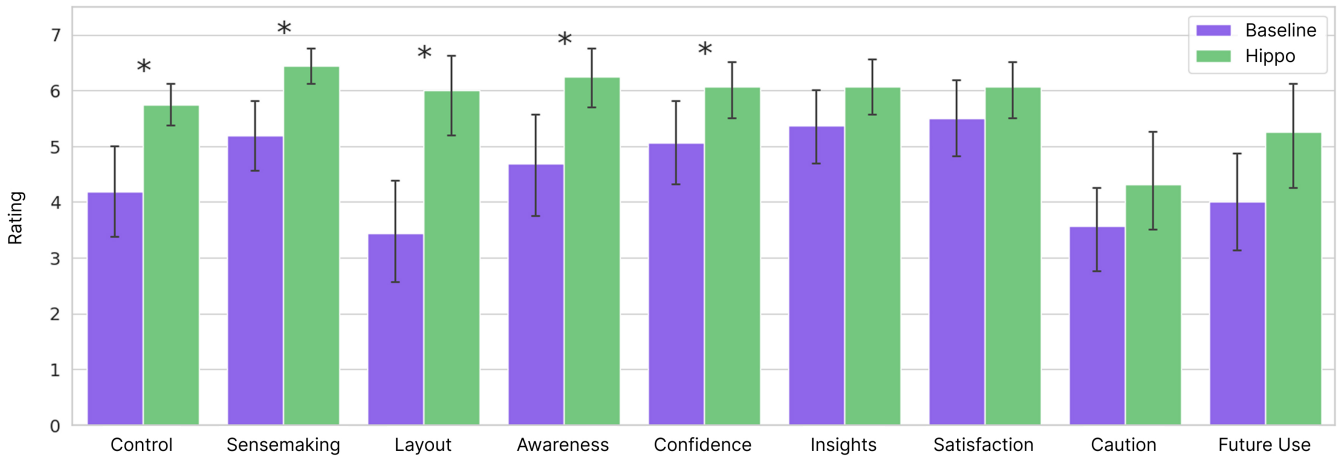

Figure 5: Participants' responses to the Likert-scale questions, contrasting the baseline and Hippo conditions. Asterisks indicate statistically significant ($p < 0.05$) differences.

4.69, $M_{\text{Hippo}} = 6.25, p = 0.012$). This improvement could be attributed to the attention that participants paid to completing the `feedback` node. Notably, P10 commented that the baseline requires "*a lot of cognitive effort to pay attention*" but they only spent less than one minute on skimming the reasoning steps in this condition. However, even though P10 also recognized they *"paid more attention"* to the tree node generation in Hippo, they found the process more engaging and felt that it helped them understand the structure of the reasoning process.

5.3.2 Response Personalization and Repurposing of Model Response. In general, participants reported feeling more confident in making a final decision in the study scenarios when using Hippo than when using the baseline system ($M_{Baseline} = 5.06, M_{\text{Hippo}} = 6.06, p = 0.049$). Many participants commented that the response after using Hippo feels personalized. For example, P8 went into more detail describing their friends in a `Feedback` node "They have toxic traits here and there. For example, they tend to gaslight a little." They were excited to see that the answer incorporated this consideration: "If they gaslight or deflect, calmly reiterate your boundaries (e.g., 'I still need to stick to my budget, so let's plan ahead')". Compared to their previous experience with reasoning models on the market, P10 commented that "*It [The output] feels like what I was expecting. So it's kind of yeah, it's kind of personalized. And I think that's cool.*" In the study, P10 had deleted a few nodes in the reasoning, which were "*just generic answers*"; in the end, they commented that the final response was "*definitely shorter to my situation*".

However, there was no significant difference between the baseline and Hippo when it comes to participants' insights ($M_{Baseline} = 5.38, M_{\text{Hippo}} = 6.06, p = 0.135$) and satisfaction over the final response ($M_{Baseline} = 5.50, M_{\text{Hippo}} = 6.06, p = 0.147$), perhaps because the model response for the baseline system was "*already very decent*," as P13 commented. What distinguished the experience was the personalized response, which made them feel that the model *"hears my voice rather than generating a generic answer."* In line with this, P6 stated:

> "*It feels like the situation is not like a wall [of text]. This is very rude to say, but when ChatGPT and DeepSeek produce things it feels like, oh, this is the most average or median output that can be created, and therefore everyone would do this. But if, after being able to provide your feedback and see what it says, it feels like this is actually tailored towards me.* "

The Likert-scale ratings also revealed no significant difference in participants' caution when reading model responses between conditions ($M_{Baseline} = 3.56, M_{\text{Hippo}} = 4.31, p = 0.267$), though participants reported being slightly more cautious with Hippo's output. In fact, we found that 7 participants felt more cautious when they read the final model output in Hippo; 2 participants were neutral; the rest of the 7 participants felt that they became less cautious after engaging with the reasoning in Hippo. For instance, compared to the baseline interface where they skimmed the reasoning, P2 reported, "*I think I spent more time working with the reasoning [with Hippo], I tend to have a better understanding of the situation already. So I trust the final response to align with my feedback.*" Similarly, P4 noted, "*the response is a good summary of the different reasonings, but I kind of feel that I don't need to see the response to make a decision for this task already. At the end of the day, it's me who is going to deal with the situation.*" The finding suggests that model reasoning can be just as valuable, if not more valuable, than the final model output itself. We further discuss the design implications of this observation in Section 7.

5.3.3 User's Interaction with Hippo. **Sparsity of Interaction on the Tree.** We observed that participants rarely made edits to the tree on the fly without the user `feedback` node. P13 explained that the content in the chain *"made a lot of sense, so I do not feel like changing anything"*. This is especially the case in the baseline condition, where only four users typed their own experiences and opinions, rather than just deleting the text or making no edits at all. Even with Hippo, P14 suggested that *"it is so easy to add or delete nodes during the reasoning process, but I rarely wanted to edit a node."*

For other participants, they indicated that they would like to make more edits after the tree completes, or even after the final response is generated. For example, P8 said "*it's good to use the graph when I revisit the reasoning, and I can just share my thought process with others if I want.*"

In the post-study interviews, participants were asked to reflect on the difference between providing feedback during the reasoning process (as implemented in our system) versus the traditional approach of iteratively prompting language models. Participants showed varying levels of preference for interactive reasoning. Some participants preferred to monitor the intermediate process. In particular, P4 mentioned that *"seeing the reasoning is so important, and I have always been wanting to edit the reasoning chain."* However, other participants suggested that they did not consider it necessary to see the reasoning steps. P9, a movie producer, commented "*if I'm having a hard time making a decision, I would want to get a few general recommendations. This process seems too logical,*" suggesting that while insights into the reasoning steps may be helpful in some situations, it may depend on users' decision styles and the contexts.

The requirement to provide feedback to the model may also lead users to doubt the model's performance, or as P9 told us: "*if the model needs to confirm with me so many times, I just felt that the model is not that good.*" The diverging opinions on interaction frequency is in line with our result that there is no significant difference between the ratings on Future Use ($M_{Baseline} = 4.00$, $M_{\text{Hippo}} = 5.25$, $p = 0.089$).

Trade-off between granular details and high-level summary. We developed Hippo to uplift the transparency of the intermediate reasoning steps and encourage users to intervene and control reasoning when possible. While many participants appreciate being more attentive to and making sense of the reasoning chains, a few participants commented on the tradeoffs between attending to the granular details and high-level summaries. P11 commented that when a user asks Hippo a question, they may not want to spend too much time focusing on the reasoning chain: "*While I read a lot of the reasoning, it might not be practical to do so in every case.*" However, most participants acknowledged that they wanted to double-check the model's reasoning if they were to ask AI for advice for themselves in the real world.

Zero participants used "Collapse the tree" function while the tree generation unfolded. Only three participants used it to collapse a subtree after the complete tree generation, primarily to revisit the overall reasoning process. When asked during post-study interviews why they did not use this feature during tree generation, participants explained they preferred maintaining full transparency of all reasoning steps on the fly. For instance, P5 mentioned that *"currently I know that I can collapse the tree after the reasoning process completes, but I kind of want to see the response quickly on the fly. Maybe you can try a more top-down approach where users can expand the tree."* This suggests a potential complementary design from the current depth-first tree traversal to a breadth-first tree traversal, which we further discuss in Section 7.

6 Case Studies

After our user study (Section 5), two participants requested to use Hippo for their real-world use cases. We followed up with the participants to explore other potential use cases and how users interact with reasoning in other decision-making scenarios. We demonstrate how they explore Hippo and design insights from these sessions. To distinguish from participants in Section 5, we refer to participants as C1 and C2. We display text that participants typed into and nodes on Hippo as "text against a light gray background".

6.1 Information Seeking

C1, a computational neuroscientist, had recently used an LLM chatbot for a question "How does hippocampus consolidate memory back to the neocortex?" for their own research studies. C1 typed and asked this question on Hippo, which generated, as C1 commented, a graphical *mindmap* of reasoning.

C1 followed the tree generation. In the first subtree, Hippo showed details about a related term "systems consolidation": "Maybe during sleep, especially slow-wave sleep, the hippocampus replays memories, which helps transfer them to the neocortex." One child node followed this topic: "There's something about sharp-wave ripples in the hippocampus during this replay." The user halted the generation and asked their follow-up question (via C from Figure 1): "Does sharp-wave ripple carry memory information?" C1 was satisfied with the response to this small point and stored this in this subtree. Later, C1 also added another follow-up question: "What are the circuits that connect hippocampus and entorhinal cortex" when the Hippo mentioned "neocortex". C1 commented that the ability to "*walk through the mindmap with the reasoning is actually very useful*", and the ability to comment and follow up with the model reasoning made them "*very engaged in the answer*" for this use case.

However, C1 wished to introduce further *control* not only to the individual nodes, but the entire tree generation process. While they acknowledged that the full reasoning chain was *"definitely more comprehensive,"* C1 commented that they "*kind of got the answer*" in the middle of the tree generation, and "*at this point, I really wanted to skip and see the [subsequent final] response, almost like a summary of what I just answered in the process.*" This indicates future design opportunities to skip subtrees where users are familiar with the subtopics and reveal the final model output promptly.

Another feedback was that the `Feedback` node interaction is not exactly what they would like to stop. For instance, Hippo stopped at a `Feedback` node for C1's clarification: "But I'm a bit fuzzy on the mechanisms of reactivation." C1 stated "*I can see why it was stopped, but I think I don't really feel like giving input here*" because this was not a point of confusion for C1. When asked about the difference between reviewing the reasoning chain and a traditional chatbot interface (e.g., ChatGPT), C1 recognized that reviewing the reasoning "*is just more engaging for this task which honestly requires a lot of attention for him*". The right or wrong answer seems less critical, compared to going through different topics in a *mindmap*.

6.2 Financial Planning

C2, a financial analyst, had recently been tasked with evaluating the potential acquisition of another company (referred to as company A). They were in the early stage of the evaluation and wanted to navigate through different aspects of this process. They typed their query "I worked at [Anonymized Company] in their corporate

development team. We are looking to acquire a company called Company X [a brief description of this company]. Help me make the case for why we should or should not acquire this company." During the interaction, we observed that C2 consistently collapsed each reasoning tree as it was generated, creating a more manageable visual hierarchy. At the `Feedback` nodes, C2 provided domain-specific expertise in response to clarify questions about Company A's market share and competitive technological positioning. This contextual information was then incorporated into subsequent branches.

Upon completing the session, C2 reflected: *"This is essentially a plan for me to consider. What this really represents is the decision-making process that happens in corporate boardrooms. What I appreciate about this approach, compared to ChatGPT or DeepSeek, is that it effectively models C-suite decision-making processes."* C2 elaborated that in large organizations, acquisition discussions involve consideration of multiple factors, and this interface allowed them to anticipate these considerations and get ahead of these ideas.

When asked to compare this experience with chatbot interfaces, C2 emphasized the value of depth over speed: *"The text alone in standard responses doesn't provide much nuance or reveal the underlying stream of consciousness. For a financial analyst, it's not about how quickly we can complete this process. I'm focused on how diligently we can review the logic behind the decision."* C2 further noted that a conventional output listing pros and cons without revealing the reasoning process would prompt immediate questions about the rationale, stating: *"If you just show the pros and cons in the final answer, I would immediately ask why is it so?"*

7 Discussion and Design Opportunities

Our work introduces user interaction with reasoning chains for LLMs. We discuss the implications of our findings as they relate to making reasoning processes transparent and paving a future design space for interaction during test-time scaling.

Improving user agency may improve perceived output quality in reasoning models. In our user study, all participants valued the transparency of reviewing the reasoning process before seeing the final model output. Reviewing the reasoning helped them understand the tradeoffs of these oftentimes subjective decision-making tasks. Having the direct entry point of interaction on Hippo made participants feel their voices are heard. Moreover, participants observed that having more direct control over the reasoning process made the final output personalized. The empirical evidence of an increased sense of control over the reasoning process suggests users benefit from engaging with the reasoning process. While work in the NLP community has emphasized the goal of test-time scaling to improve the model output [66], the observations in our study challenge recommendations [3, 51] that suggest concealing intermediate reasoning processes from users or limiting user control over the reasoning process.

Re-purposing model output after interactive reasoning. We found mixed results on participants' caution over the final model output. While prior work in XAI suggests that cognitive forcing functions may reduce overreliance on the model output, our findings add nuances. For high-stakes decision-making tasks, users may consult LLMs for tradeoffs but may defer to themselves to make such decisions. In fact, as P2 put it in Section 5.3.2, users may just gain "*a better understanding of the situation already*" by going through the reasoning process. While test-time scaling is now primarily viewed as a means to improve model responses [3, 66], in the context of decision-making tasks, the final response could be repurposed as a personalized summary, rather than the main objective that models typically produce. This finding suggests a potential shift in how we conceptualize the "output" of AI systems—from the final objective to supportive reasoning artifacts that enhance users' decision-making.

Tensions in designing for interactive reasoning. Our study also revealed potential difficulties in how users engage with LLM reasoning. Participants found that the long presentation of reasoning in the baseline created significant cognitive barriers that prevented meaningful engagement, despite participants' expressed interest in understanding the model's thought process. Hippo alleviated some of these problems by transforming the "wall of text" into an interactive tree representation. We note that participants were more aware of the assumptions from the reasoning by interacting with Hippo; after all, making an informed decision requires navigating nuanced trade-offs. However, the tree representation—or even showing any reasoning at all—might be unnecessary for simpler or low-stakes queries. Indeed, some participants wished to bypass the reasoning process altogether after providing some feedback (e.g., C1). This surfaces a design opportunity to calibrate the display of and interaction with reasoning based on task complexity. This might involve adaptive interfaces that present abbreviated reasoning for routine decisions while reserving comprehensive reasoning chains for ones that demand more complex tradeoffs.

Tradeoffs of the depth-first and breadth-first tree structure. Hippo was motivated by projects in the UIST community that use a node-link diagram for information sense-making [25, 70]. Our findings (Section 5.3.1) of improved sense-making corroborate prior work in the context of reviewing long reasoning during test-time scaling. Meanwhile, we note a potential design opportunity to edit AI reasoning on the fly. Many participants suggested new features to view reasoning at multiple levels of abstraction—first seeing broad conceptual frameworks before selectively exploring specific details of interest (i.e., a *breadth-first* tree traversal). Contrasting this was the Hippo design which followed a *depth-first* tree traversal that followed the original trajectory of the reasoning chains. However, as some participants suggested, reasoning involves complex networks of interconnected concepts, premises, and inferences, rather than a purely sequential or linear fashion in current interfaces (including in Hippo). The current sequential reasoning paradigm from the reasoning model and the exploration based on hierarchical order reflects a *mismatch between current interface paradigms and the nature of reasoning* [27]. Future designs could better support interactive reasoning by adopting visualization approaches that explicitly represent *networked* relationships. This could involve interfaces that support *both breadth-first exploration for context and depth-first exploration for details*, with interactive capabilities to expand or collapse reasoning branches according to user interest. Such approaches could bridge the gap between the linear presentation of current interfaces and the more networked structure of human reasoning.

Varied Engagement Preferences in LLM Reasoning. We found significant variation in how participants wished to engage

with reasoning processes. On the one hand, some participants wished to reduce the engagement level, stating that "*if the model needs to confirm with me so many times, I just felt that the model is not that good.*" [P9] Other participants wish to include even more interaction, such as eliciting users' confirmation when a node branches off. This suggests that future design with model reasoning (e.g., [43]) should require considerations such as user's existing level of AI reliance and task complexity. For instance, participants noted that the tree visualization was more valuable for analytical questions than for creative or simple factual tasks. Some participants might prioritize efficiency, while others value learning from the reasoning process itself. These findings suggest design opportunities for interactive reasoning that accommodates varying levels of engagement across different contexts and decision-making styles [46]. Potential designs that support this variance may range from fully automated reasoning with minimal visibility to highly interactive approaches in which users actively shape the reasoning trajectory.

8 Limitations and Future Work

Limitations of this work are the relatively limited sample size (N=16) as well as participants' relatively high familiarity with LLM models, which may limit the generalizability of our findings to novice populations. We attempted to alleviate this limitation by recruiting participants from diverse backgrounds ranging from undergraduate students to working professionals in finance, computer science, and art. While we used two daily dilemma scenarios to increase the stakes of the task and allow fair comparison between the two conditions in the study, decision-making in knowledge-intensive scenarios (e.g., coding [86] or medical diagnosis [72]) might require information from beyond just model reasoning.

Another limitation is that our current approach does not systematically analyze the model behaviors given users' feedback. In fact, a recent study [10] of reasoning models during test-time scaling without human feedback shows that CoTs may not faithfully represent a model's actual reasoning process to answer math and coding problems. While our paper does not claim that user feedback led to more personalized or accurate reasoning chains or final model outputs, participants observed that their inputs were incorporated into the final outputs, which demonstrated several benefits from the end user perspective. We also acknowledge that the *Link* operator matches the sentences in the model response and the content within each reasoning node based on semantic similarity; however, this linkage does not indicate a causal relationship internal to the model. Future work may consider investigating more robust approaches to derive this causal relationship for further explainability.

We acknowledge that Hippo may incur misuse for ethical consideration. We recognize that LLMs, such as `gpt-4o` and `DeepSeek-R1` can hallucinate and generate false information that may affect daily decision makings. Users might also mistakenly steer model output by providing existing biased or malicious feedback, leading to harmful result but became more confident in the end [62]. To remedy this, we cautioned users for such risks in our system and study. We also chose daily dilemma tasks in the user study rather than topics that can be highly controversial, such as political disagreement [13].

9 Conclusion

In this paper, we introduced Hippo, a system that instantiates *interactive reasoning*, an approach that visualizes the LLM reasoning steps via test-time scaling and allows users to make sense of, and control the reasoning before reaching the model's final output. We evaluated Hippo through a user study with 16 participants from diverse occupational backgrounds, as well as two case studies to explore how users leverage interactive reasoning in decision-making tasks. Results showed that Hippo increased the sense of control, sense-making, and assumption in the model output compared to a baseline system. We also observed participants' repurposing of model responses after engaging with the reasoning steps. We discuss practical implications for future adaptive designs that support interaction with a model's reasoning steps. Overall, our contributions set the stage for new interactive paradigms for test-time scaling not just for model response quality but also to improve user control through human agency.

Acknowledgments

We thank our participants and the anonymous reviewers for their valuable feedback. We also thank Faeze Brahman, Liwei Jiang, Ruotong Wang, Katelyn Mei, Alice Gao for their helpful suggestions. This work was funded by the National Science Foundation as part of the awards ER2-2315937 and IBM PhD Fellowship.

References

[1] Tyler Angert, Miroslav Suzara, Jenny Han, Christopher Pondoc, and Hariharan Subramonyam. 2023. Spellburst: A Node-based Interface for Exploratory Creative Coding with Natural Language Prompts. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) *(UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 100, 22 pages. doi:10.1145/3586183.3606719

[2] Ian Arawjo, Chelse Swoopes, Priyan Vaithilingam, Martin Wattenberg, and Elena L. Glassman. 2024. ChainForge: A Visual Toolkit for Prompt Engineering and LLM Hypothesis Testing. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 304, 18 pages. doi:10.1145/3613904.3642016

[3] Bowen Baker, Joost Huizinga, Leo Gao, Zehao Dou, Melody Y. Guan, Aleksander Madry, Wojciech Zaremba, Jakub Pachocki, and David Farhi. 2025. Monitoring Reasoning Models for Misbehavior and the Risks of Promoting Obfuscation. arXiv:2503.11926 [cs.AI] https://arxiv.org/abs/2503.11926

[4] Jeff Baker, Donald Jones, and Jim Burkman. 2009. Using visual representations of data to enhance sensemaking in data exploration tasks. *Journal of the Association for Information Systems* 10, 7 (2009), 2.

[5] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 81, 16 pages. doi:10.1145/3411764.3445717

[6] Jessica Y Bo, Sophia Wan, and Ashton Anderson. 2025. To Rely or Not to Rely? Evaluating Interventions for Appropriate Reliance on Large Language Models. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 905, 23 pages. doi:10.1145/3706598.3714097

[7] Virginia Braun and Victoria Clarke and. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. doi:10.1191/1478088706qp063oa arXiv:https://www.tandfonline.com/doi/pdf/10.1191/1478088706qp063oa

[8] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 188 (April 2021), 21 pages. doi:10.1145/3449287

[9] Boqi Chen, Fandi Yi, and Dániel Varró. 2023. Prompting or fine-tuning? a comparative study of large language models for taxonomy construction. In *2023 ACM/IEEE International Conference on Model Driven Engineering Languages and*

Systems Companion (MODELS-C). IEEE, 588–596.

[10] Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, Vlad Mikulik, Samuel R. Bowman, Jan Leike, Jared Kaplan, and Ethan Perez. 2025. Reasoning Models Don't Always Say What They Think. arXiv:2505.05410 [cs.CL] https://arxiv.org/abs/2505.05410

[11] Yu Ying Chiu, Liwei Jiang, and Yejin Choi. 2024. Dailydilemmas: Revealing value preferences of llms with quandaries of daily life. *arXiv preprint arXiv:2410.02683* (2024).

[12] Hai Dang, Sven Goller, Florian Lehmann, and Daniel Buschek. 2023. Choice Over Control: How Users Write with Large Language Models using Diegetic and Non-Diegetic Prompting. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 408, 17 pages. doi:10.1145/3544548.3580969

[13] Jillian Fisher, Shangbin Feng, Robert Aron, Thomas Richardson, Yejin Choi, Daniel W. Fisher, Jennifer Pan, Yulia Tsvetkov, and Katharina Reinecke. 2025. Biased AI can Influence Political Decision-Making. arXiv:2410.06415 [cs.HC] https://arxiv.org/abs/2410.06415

[14] Austin J Freeley. 2009. *Argumentation and debate: Critical thinking for reasoned decision making*.

[15] Katy Ilonka Gero, Chelse Swoopes, Ziwei Gu, Jonathan K. Kummerfeld, and Elena L. Glassman. 2024. Supporting Sensemaking of Large Language Model Outputs at Scale. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 838, 21 pages. doi:10.1145/3613904.3642139

[16] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948 [cs.CL] https://arxiv.org/abs/2501.12948

[17] Mary E Guy. 1990. *Ethical decision making in everyday work situations*. Bloomsbury Publishing.

[18] Zeyu He, Chieh-Yang Huang, Chien-Kuang Cornelia Ding, Shaurya Rohatgi, and Ting-Hao Kenneth Huang. 2024. If in a Crowdsourced Data Annotation Pipeline, a GPT-4. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 1040, 25 pages. doi:10.1145/3613904.3642834

[19] Jeffrey Heer. 2019. Agency plus automation: Designing artificial intelligence into interactive systems. *Proceedings of the National Academy of Sciences* 116, 6 (2019), 1844–1850.

[20] Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2020. Aligning ai with shared human values. *arXiv preprint arXiv:2008.02275* (2020).

[21] Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. 2021. Surface form competition: Why the highest probability answer isn't always right. *arXiv preprint arXiv:2104.08315* (2021).

[22] Md Naimul Hoque, Tasfia Mashiat, Bhavya Ghai, Cecilia D. Shelton, Fanny Chevalier, Kari Kraus, and Niklas Elmqvist. 2024. The HaLLMark Effect: Supporting Provenance and Transparent Use of Large Language Models in Writing with Interactive Visualization. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 1045, 15 pages. doi:10.1145/3613904.3641895

[23] Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Pittsburgh, Pennsylvania, USA) *(CHI '99)*. Association for Computing Machinery, New York, NY, USA, 159–166. doi:10.1145/302979.303030

[24] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. OpenAI o1 System Card. arXiv:2412.16720 [cs.AI] https://arxiv.org/abs/2412.16720

[25] Peiling Jiang, Jude Rayan, Steven P. Dow, and Haijun Xia. 2023. Graphologue: Exploring Large Language Model Responses with Interactive Diagrams. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) *(UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 3, 20 pages. doi:10.1145/3586183.3606737

[26] Zhuohang Jiang, Pangjing Wu, Ziran Liang, Peter Q Chen, Xu Yuan, Ye Jia, Jiancheng Tu, Chen Li, Peter HF Ng, and Qing Li. 2025. HiBench: Benchmarking LLMs Capability on Hierarchical Structure Reasoning. *arXiv preprint arXiv:2503.00912* (2025).

[27] Philip N Johnson-Laird, Sangeet S Khemlani, and Geoffrey P Goodwin. 2015. Logic, probability, and human reasoning. *Trends in cognitive sciences* 19, 4 (2015), 201–214.

[28] Minsuk Kahng, Ian Tenney, Mahima Pushkarna, Michael Xieyang Liu, James Wexler, Emily Reif, Krystal Kallarackal, Minsuk Chang, Michael Terry, and Lucas Dixon. 2024. LLM Comparator: Visual Analytics for Side-by-Side Evaluation of Large Language Models. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, Article 216, 7 pages. doi:10.1145/3613905.3650755

[29] Hita Kambhamettu, Jamie Flores, and Andrew Head. 2025. Traceable Texts and Their Effects: A Study of Summary-Source Links in AI-Generated Summaries. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, Article 538, 7 pages. doi:10.1145/3706599.3719830

[30] Majeed Kazemitabaar, Jack Williams, Ian Drosos, Tovi Grossman, Austin Zachary Henley, Carina Negreanu, and Advait Sarkar. 2024. Improving Steering and Verification in AI-Assisted Data Analysis with Interactive Task Decomposition. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (Pittsburgh, PA, USA) *(UIST '24)*. Association for Computing Machinery, New York, NY, USA, Article 92, 19 pages. doi:10.1145/3654777.3676345

[31] Sunnie S. Y. Kim, Q. Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. 2024. "I'm Not Sure, But...": Examining the Impact of Large Language Models' Uncertainty Expression on User Reliance and Trust. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 822–835. doi:10.1145/3630106.3658941

[32] Sunnie S. Y. Kim, Jennifer Wortman Vaughan, Q. Vera Liao, Tania Lombrozo, and Olga Russakovsky. 2025. Fostering Appropriate Reliance on Large Language Models: The Role of Explanations, Sources, and Inconsistencies. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 420, 19 pages. doi:10.1145/3706598.3714020

[33] Sunnie S. Y. Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández. 2023. "Help Me Help the AI": Understanding How Explainability Can Support Human-AI Interaction. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 250, 17 pages. doi:10.1145/3544548.3581001

[34] Tae Soo Kim, Yoonjoo Lee, Jamin Shin, Young-Ho Kim, and Juho Kim. 2024. EvalLM: Interactive Evaluation of Large Language Model Prompts on User-Defined Criteria. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 306, 21 pages. doi:10.1145/3613904.3642216

[35] Yoonsu Kim, Jueon Lee, Seoyoung Kim, Jaehyuk Park, and Juho Kim. 2024. Understanding Users' Dissatisfaction with ChatGPT Responses: Types, Resolving Tactics, and the Effect of Knowledge Level. In *Proceedings of the 29th International Conference on Intelligent User Interfaces* (Greenville, SC, USA) *(IUI '24)*. Association for Computing Machinery, New York, NY, USA, 385–404. doi:10.1145/3640543.3645148

[36] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large Language Models are Zero-Shot Reasoners. arXiv:2205.11916 [cs.CL] https://arxiv.org/abs/2205.11916

[37] Michelle S. Lam, Janice Teoh, James A. Landay, Jeffrey Heer, and Michael S. Bernstein. 2024. Concept Induction: Analyzing Unstructured Text with High-Level Concepts Using LLooM. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 766, 28 pages. doi:10.1145/3613904.3642830

[38] Hao-Ping (Hank) Lee, Advait Sarkar, Lev Tankelevitch, Ian Drosos, Sean Rintel, Richard Banks, and Nicholas Wilson. 2025. The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 1121, 22 pages. doi:10.1145/3706598.3713778

[39] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 7871–7880. doi:10.18653/v1/2020.acl-main.703

[40] Diya Li, Yue Zhao, Zhifang Wang, Calvin Jung, and Zhe Zhang. 2024. Large Language Model-Driven Structured Output: A Comprehensive Benchmark and Spatial Data Generation Framework. *ISPRS International Journal of Geo-Information* (2024). https://api.semanticscholar.org/CorpusID:273976782

[41] Jiacheng Liu, Andrew Cohen, Ramakanth Pasunuru, Yejin Choi, Hannaneh Hajishirzi, and Asli Celikyilmaz. 2024. Don't throw away your value model! Generating more preferable text with Value-Guided Monte-Carlo Tree Search decoding. arXiv:2309.15028 [cs.CL] https://arxiv.org/abs/2309.15028

[42] Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics* 12 (2024), 157–173.

[43] Xingyu Bruce Liu, Haijun Xia, and Xiang Anthony Chen. 2025. Interacting with Thoughtful AI. *arXiv preprint arXiv:2502.18676* (2025).

[44] Yang Liu, Alex Kale, Tim Althoff, and Jeffrey Heer. 2021. Boba: Authoring and Visualizing Multiverse Analyses. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (2021), 1753–1763. doi:10.1109/TVCG.2020.3028985

[45] Damien Masson, Sylvain Malacria, Géry Casiez, and Daniel Vogel. 2024. DirectGPT: A Direct Manipulation Interface to Interact with Large Language Models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 975, 16 pages. doi:10.1145/3613904.3642462

[46] Katelyn Xiaoying Mei, Rock Yuren Pang, Alex Lyford, Lucy Lu Wang, and Katharina Reinecke. 2025. Passing the Buck to AI: How Individuals' Decision-Making Patterns Affect Reliance on AI. arXiv:2505.01537 [cs.HC] https://arxiv.org/abs/2505.01537

[47] Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work? *ArXiv* abs/2202.12837 (2022). https://api.semanticscholar.org/CorpusID:247155069

[48] Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. 2024. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. *arXiv preprint arXiv:2410.05229* (2024).

[49] Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. arXiv:2501.19393 [cs.CL] https://arxiv.org/abs/2501.19393

[50] Richard Nordquist. 2019. What is deductive reasoning? https://www.thoughtco.com/deduction-logic-and-rhetoric-1690422

[51] OpenAI. 2025. Detecting Misbehavior in Frontier Reasoning Models. https://openai.com/index/chain-of-thought-monitoring/. Accessed Apr 08, 2025.

[52] Siru Ouyang, Shuohang Wang, Yang Liu, Ming Zhong, Yizhu Jiao, Dan Iter, Reid Pryzant, Chenguang Zhu, Heng Ji, and Jiawei Han. 2023. The Shifted and The Overlooked: A Task-oriented Investigation of User-GPT Interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 2375–2393. doi:10.18653/v1/2023.emnlp-main.146

[53] Rock Yuren Pang, Hope Schroeder, Kynnedy Simone Smith, Solon Barocas, Ziang Xiao, Emily Tseng, and Danielle Bragg. 2025. Understanding the LLM-ification of CHI: Unpacking the Impact of LLMs at CHI through a Systematic Literature Review. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 456, 20 pages. doi:10.1145/3706598.3713726

[54] Joon Sung Park, Rick Barber, Alex Kirlik, and Karrie Karahalios. 2019. A slow algorithm improves users' assessments of the algorithm's accuracy. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–15.

[55] Andreas Peldszus and Manfred Stede. 2013. From argument diagrams to argumentation mining in texts: A survey. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)* 7, 1 (2013), 1–31.

[56] Vincent Perot, Kai Kang, Florian Luisier, Guolong Su, Xiaoyu Sun, Ramya Sree Boppana, Zilong Wang, Zifeng Wang, Jiaqi Mu, Hao Zhang, Chen-Yu Lee, and Nan Hua. 2024. LMDX: Language Model-based Document Information Extraction and Localization. In *Findings of the Association for Computational Linguistics: ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 15140–15168. doi:10.18653/v1/2024.findings-acl.899

[57] Kevin Pu, K. J. Kevin Feng, Tovi Grossman, Tom Hope, Bhavana Dalvi Mishra, Matt Latzke, Jonathan Bragg, Joseph Chee Chang, and Pao Siangliulue. 2025. IdeaSynth: Iterative Research Idea Development Through Evolving and Composing Idea Facets with Literature-Grounded Feedback. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 145, 31 pages. doi:10.1145/3706598.3714057

[58] Sebastian Raschka. 2025. Understanding reasoning llms. https://magazine.sebastianraschka.com/p/understanding-reasoning-llms

[59] Chris Reed, Douglas Walton, and Fabrizio Macagno. 2007. Argument diagramming in logic, law and artificial intelligence. *The Knowledge Engineering Review* 22, 1 (2007), 87–109.

[60] Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. 2022. Multitask Prompted Training Enables Zero-Shot Task Generalization. In *International Conference on Learning Representations*. https://openreview.net/forum?id=9Vrb9D0WI4

[61] Max Schemmer, Niklas Kuehl, Carina Benz, Andrea Bartos, and Gerhard Satzger. 2023. Appropriate Reliance on AI Advice: Conceptualization and the Effect of Explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) *(IUI '23)*. Association for Computing Machinery, New York, NY, USA, 410–422. doi:10.1145/3581641.3584066

[62] Nikhil Sharma, Q. Vera Liao, and Ziang Xiao. 2024. Generative Echo Chamber? Effect of LLM-Powered Search Systems on Diverse Information Seeking. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 1033, 17 pages. doi:10.1145/3613904.3642459

[63] Ben Shneiderman. 1983. Direct manipulation: A step beyond programming languages. *Computer* 16, 08 (1983), 57–69.

[64] Ben Shneiderman. 2020. Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy. *International Journal of Human–Computer Interaction* 36 (2020), 495 – 504. https://api.semanticscholar.org/CorpusID:211259461

[65] Connor Shorten, Charles Pierse, Thomas Benjamin Smith, Erika Cardenas, Akanksha Sharma, John Trengrove, and Bob van Luijt. 2024. StructuredRAG: JSON Response Formatting with Large Language Models. *ArXiv* abs/2408.11061 (2024). https://api.semanticscholar.org/CorpusID:271916259

[66] Charlie Victor Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2025. Scaling LLM Test-Time Compute Optimally Can be More Effective than Scaling Parameters for Reasoning. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=4FWAwZtd2n

[67] Paul M Sniderman and Sean M Theriault. 2004. The structure of political argument and the logic of issue framing. *Studies in public opinion: Attitudes, nonattitudes, measurement error, and change* 3, 03 (2004), 133–65.

[68] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models.

[69] Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics* 43, 3 (2017), 619–659.

[70] Sangho Suh, Bryan Min, Srishti Palani, and Haijun Xia. 2023. Sensecape: Enabling Multilevel Exploration and Sensemaking with Large Language Models. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) *(UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 1, 18 pages. doi:10.1145/3586183.3606756

[71] Siddharth Swaroop, Zana Buçinca, Krzysztof Z. Gajos, and Finale Doshi-Velez. 2024. Accuracy-Time Tradeoffs in AI-Assisted Decision Making under Time Pressure. In *Proceedings of the 29th International Conference on Intelligent User Interfaces* (Greenville, SC, USA) *(IUI '24)*. Association for Computing Machinery, New York, NY, USA, 138–154. doi:10.1145/3640543.3645206

[72] Peter Szolovits, Ramesh S Patil, and William B Schwartz. 1988. Artificial intelligence in medical diagnosis. *Annals of internal medicine* 108, 1 (1988), 80–87.

[73] Together AI. 2025. Together API. https://api.together.xyz/. Accessed April 08, 2025.

[74] Douglas N Walton. 1990. What is reasoning? What is an argument? *The journal of Philosophy* 87, 8 (1990), 399–419.

[75] Douglas N Walton and Lynn M Batten. 1984. Games, graphs and circular arguments. *Logique et Analyse* 27, 106 (1984), 133–164.

[76] Heng Wang, Shangbin Feng, Tianxing He, Zhaoxuan Tan, Xiaochuang Han, and Yulia Tsvetkov. 2023. Can language models solve graph problems in natural language? *Advances in Neural Information Processing Systems* 36 (2023), 30840–30861.

[77] Xinru Wang and Ming Yin. 2021. Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. In *Proceedings of the 26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) *(IUI '21)*. Association for Computing Machinery, New York, NY, USA, 318–328. doi:10.1145/3397481.3450650

[78] Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, et al. 2022. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. *arXiv preprint arXiv:2204.07705* (2022).

[79] Zijie J. Wang, Chinmay Kulkarni, Lauren Wilcox, Michael Terry, and Michael Madaio. 2024. Farsight: Fostering Responsible AI Awareness During AI Application Prototyping. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 976, 40 pages. doi:10.1145/3613904.3642335

[80] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.

[81] Tongshuang Wu, Ellen Jiang, Aaron Donsbach, Jeff Gray, Alejandra Molina, Michael Terry, and Carrie J Cai. 2022. PromptChainer: Chaining Large Language Model Prompts through Visual Programming. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA)

(CHI EA '22). Association for Computing Machinery, New York, NY, USA, Article 359, 10 pages. doi:10.1145/3491101.3519729

[82] Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022. AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 385, 22 pages. doi:10.1145/3491102.3517582

[83] Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. 2024. Inference scaling laws: An empirical analysis of compute-optimal inference for problem-solving with language models. *arXiv preprint arXiv:2408.00724* (2024).

[84] Yuxi Xie, Anirudh Goyal, Wenyue Zheng, Min-Yen Kan, Timothy P Lillicrap, Kenji Kawaguchi, and Michael Shieh. 2024. Monte carlo tree search boosts reasoning via iterative preference learning. *arXiv preprint arXiv:2405.00451* (2024).

[85] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik R Narasimhan. 2023. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. In *Thirty-seventh Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=5Xc1ecxO1h

[86] Ryan Yen, Jiawen Stefanie Zhu, Sangho Suh, Haijun Xia, and Jian Zhao. 2024. CoLadder: Manipulating Code Generation via Multi-Level Blocks. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (Pittsburgh, PA, USA) *(UIST '24)*. Association for Computing Machinery, New York, NY, USA, Article 11, 20 pages. doi:10.1145/3654777.3676357

[87] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the Effect of Accuracy on Trust in Machine Learning Models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3290605.3300509

[88] Fei Yu, Hongbo Zhang, Prayag Tiwari, and Benyou Wang. 2024. Natural Language Reasoning, A Survey. *ACM Comput. Surv.* 56, 12, Article 304 (Oct. 2024), 39 pages. doi:10.1145/3664194

[89] Fei Yu, Hongbo Zhang, Prayag Tiwari, and Benyou Wang. 2024. Natural language reasoning, a survey. *Comput. Surveys* 56, 12 (2024), 1–39.

[90] J.D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 437, 21 pages. doi:10.1145/3544548.3581388

[91] Amy X. Zhang, Lea Verou, and David Karger. 2017. Wikum: Bridging Discussion Forums and Wikis Using Recursive Summarization. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 2082–2096. doi:10.1145/2998181.2998235

[92] Qiyuan Zhang, Fuyuan Lyu, Zexu Sun, Lei Wang, Weixu Zhang, Wenyue Hua, Haolun Wu, Zhihan Guo, Yufei Wang, Niklas Muennighoff, Irwin King, Xue Liu, and Chen Ma. 2025. A Survey on Test-Time Scaling in Large Language Models: What, How, Where, and How Well? arXiv:2503.24235 [cs.CL] https://arxiv.org/abs/2503.24235

[93] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 295–305. doi:10.1145/3351095.3372852

[94] Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. Wildchat: 1m chatgpt interaction logs in the wild. *arXiv preprint arXiv:2405.01470* (2024).

[95] Zhaocheng Zhu, Yuan Xue, Xinyun Chen, Denny Zhou, Jian Tang, Dale Schuurmans, and Hanjun Dai. 2023. Large language models can learn rules. *arXiv preprint arXiv:2310.07064* (2023).

A An example of DeepSeek-R1 response with the reasoning steps

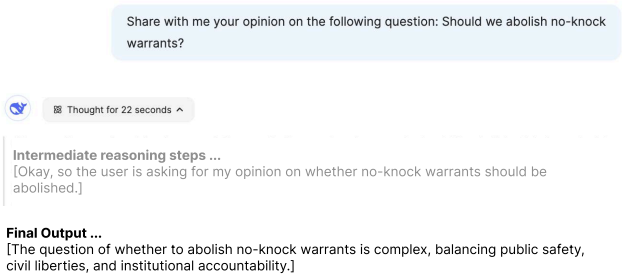

Figure 6: An example of the current DeepSeek-R1 platform that shows the reasoning steps. The model thought for 22 seconds with over 600 words in the reasoning.

B The System Prompts for Operators in the Tree Generation Pipeline

B.1 The Structure Operator

This prompt structures the original chain-of-thoughts (CoTs) reasoning into sub-topics. Before applying this operator, we grouped the thoughts first (❶ Figure 3) into manageable topics, in line with prior work to leverage LLMs to aggregate concepts [37]. In the project, we observed that CoTs during test-time scaling often lead to many paragraphs, which made the tree very sparse/shallow without this step. Below are the prompts for both the *Structure* operator and the group thought pre-processing.

B.1.1 Structure. `You are a helpful assistant that *only* tags the chain of thought for a given text.`

`Rule 1: Use <topic>...</topic> to indicate a major new area of thought.`

`Rule 2: Use <branch>...</branch> to indicate a subtopic extending from a previous point.`

`Rule 3: Nesting Structure: 1. All content must be contained within tags, with no unmarked text. 2. <topic> tags should only appear at the top level. 3. <branch> tags can nest inside <topic> tags or other <branch> tags. 4. Each <branch> must begin with at least one complete sentence before any nested branches. Please follow the rules strictly.`

`Example:`

```
<Topic> Okay, the user is asking where they should travel during spring break. Hmm, I need to consider different types of destinations to cover various interests.
</Topic>
<Topic> First, I should consider different destinations. I should list options that cater to different preferences.
</Topic>
<Branch>
   <Branch>
      For beach destinations, places like Cancun, Miami, or the Caribbean come to mind.
      <Branch>
```

These spots are popular for spring break because they offer warm weather, beaches, and vibrant nightlife.
</Branch>
</Branch>
<Branch>
But maybe I should also include some less crowded beaches for those who prefer a quieter time.
<Branch>
Costa Rica is known for eco-tourism, rainforests, and activities like zip-lining.
</Branch>
<Branch>
Hawaii is another option with hiking and volcanoes.
</Branch>
</Branch>
<Branch>
Cities with cultural attractions could be another category: places like Paris, Kyoto, or Barcelona. Oh, Kyoto in spring would have cherry blossoms, that's a big plus.
</Branch>
</Branch>
<Branch>
Wait, what is the user's budget? I should include some budget-friendly options too.
</Branch> [More of this example]

B.1.2 Group Thoughts. I have this monologue, representing my reasoning for the query: ${query}. Structure this text into high-level themes.

Rule 1: Keep the exact text from the input text!! Use the same words. Each theme should surround a high-level idea.

Rule 2: New line to separate each theme. The output should be a direct division of the input text into themes under 8 paragraphs.

Rule 3: The output should not include anything beyond the origin input text, or any summary. I'd like to the exact same text from the input.

Input: ${reasoning.replaceAll("\n", "")}

Output:

B.2 The Clarify Operator

You are a helpful assistant that *only* tag the chain of thought (which was generated by a model) for a given text. The goal is for users to help clarify the uncertain or incorrect assumptions in the input reasoning chain. You use <user></user> to tag such text.

Rule 1: Identify places where user input would be valuable (uncertainty, preferences, personal experiences)

Rule 2: Sentences like "I don't know X" that the reasoning chain is unsure about the situation.

Rule 3: Preserve the original text from the input and only add the <user> tag to the sentences that need clarification.

Rule 4: A good user should tag an uncertain question or a situation so that a user can easily give their feedback or context.

Rule 5: Do not tag questions that are answered in the reasoning chain later in the text.

Rule 6: The user tag should only appear between <branch> tags; no user tag between <topic> tags allowed.

Bad example: <branch>Hmmm. Let me think. </branch>

Good example: <branch><user>But then, how to enforce that? It's tricky because everyone has different schedules. What rules would be reasonable to create and enforce in this situation?</user></branch>

<Branch>
<user>Wait, what is the user's budget? I should include some budget-friendly options too.**</user>**
</Branch> [More of this example]

B.3 The Link Operator

Given the following premises (reasoning nodes) and hypotheses (response paragraphs), determine the entailment relationship between them.

PREMISES: [
{"id": $node_1_id, "content": $node_1_content},
{"id": $node_2_id, "content": $node_2_content}, ...]
HYPOTHESES: [
{"id": $response_1_id, "content": $response_1_content},
{"id": $response_2_id, "content": $response_2_content}, ...]

For each hypothesis (response paragraph), identify the premise that most strongly entails or supports it. Consider the semantic and logical relationship between each premise-hypothesis pair.

Return your analysis as a valid JSON array with objects containing:

```
{
    "hypothesis_id": [response ID number],
    "entailing_premise": {
    "premise_id": [most relevant node ID],
    "entailment_strength": [confidence score between 0 and 1]
    }
}
```

B.4 Other Prompts

Hippo also has a summarization feature to collapse the (sub)-trees. This operator aggregates the subtree nodes into a low-level context, then performs a summarization with GPT-4o. Note that when users expand the summarized subtree, the child nodes remain the same and are not regenerated.

Given the context, please summarize these thoughts into a paragraph of summary under 60 words.

Content: ${subtree_context} One sentence summary: